32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

An Accurate and Interpretable Model for BCCT.core

Hélder P. Oliveira, *Student Member*, *IEEE*, André Magalhães, Maria J. Cardoso,
Jaime S. Cardoso, *Member*, *IEEE*

***Abstract*— Breast Cancer Conservative Treatment (BCCT) is considered nowadays to be the most widespread form of locorregional breast cancer treatment. However, aesthetic results are heterogeneous and difficult to evaluate in a standardized way. The limited reproducibility of subjective aesthetic evaluation in BCCT motivated the research towards objective methods. A recent computer system (BCCT.core) was developed to objectively and automatically evaluate the aesthetic result of BCCT. The system is centered on a support vector machine (SVM) classifier with a radial basis function (RBF) used to predict the overall cosmetic result from features computed on a digital photograph of the patient. However, this classifier is not ideal for the interpretation of the factors being used in the prediction. Therefore, an often suggested improvement is the interpretability of the model being used to assess the overall aesthetic result.**

In the current work we investigate the accuracy of different interpretable methods against the model currently deployed in the BCCT.core software. We compare the performance of decision trees and linear classifiers with the RBF SVM currently in BCCT.core. In the experimental study, these interpretable models shown a similar accuracy to the currently used RBF SVM, suggesting that the later can be replaced without sacrificing the performance of the BCCT.core.

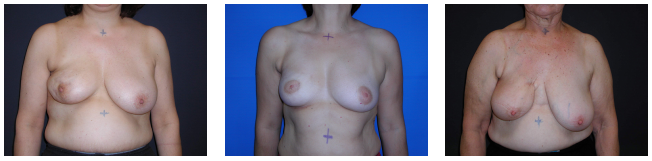
I. Introduction

Breast cancer is the most common cancer to affect women in Europe and as 10-year survival from the disease now exceeds 80%, many women are expected to live a long time with the aesthetic consequences of their treatment. Therefore, a good aesthetic outcome is an important endpoint of breast cancer treatment, being closely related to psychosocial recovery and quality of life. The importance of a good aesthetic outcome is well recognized by experts in this field although it is known that this is often not achieved. In breast-conserving surgery for example, approximately 33% of women will have a fair or poor aesthetic outcome [1], [2].

A significant obstacle in auditing this problem and evaluating techniques for improving it has been the absence of a standard method for measuring the aesthetic outcome. The most commonly used methods until recently, involved subjective assessment by an expert panel. Initial objective methods consisted on the comparison between the two breasts with simple measurements marked directly in patients or in photographs [3], [4]. Trying to overcome the sense that objective asymmetry measurements were insufficient, other groups proposed the sum of the individual scores of subjective and objective individual indices [5]. Even with these additions all available methods were subject to significant intra-observer and inter-observer variability. There was a need to replace or enhance human expert evaluation of the aesthetic results of BCCT with a validated objective tool. This tool should be easy to use and highly reproducible.

More recently, a computer-aided medical system was developed to objectively and automatically perform the aesthetic evaluation of BCCT [6]. This system, named BCCT.core, aims to overcome the acute shortage of such software systems and exploit the unique ability of computational methods to provide an effective and easy to use tool for one important outcome of breast cancer patient care. BCCT.core is an automatic system capable of objectively evaluating the overall aesthetic result of BCCT.

The development of BCCT.core entailed the automatic extraction of several features from the patient's photographs (Fig. 1), capturing some of the factors considered to have impact on the overall cosmetic result: breast asymmetry, skin colour changes due to the radiotherapy treatment and surgical scar appearance. In a second phase, a SVM classifier was trained to predict the overall cosmetic result from the recorded features [6].



(a) (b) (c)

Fig. 1. Typical photographs.

Although BCCT.core is currently being used by many international groups in prospective studies, one often suggested improvement is the interpretability of the model being used to assess the overall aesthetic result. Although SVMs have proven to be very useful in machine learning, there is a significant drop of understandability of the learned hypothesis, especially when using nonlinear kernels, as is the case of the RBF kernel in BCCT.core.

In the current work we investigate the improvement of BCCT.core, by comparing the performance of different interpretable methods against the currently deployed in

First author was supported for a PhD grant by Fundação para a Ciência e Tecnologia (FCT) with reference SFRH/BD/43772/2008. This work was also partially funded by FCT through project PTDC/EIA/64914/2006.

H. P. Oliveira and J. S. Cardoso are with INESC Porto, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 378, 4200-465 Porto, Portugal. helder.oliveira@fe.up.pt, jaime.cardoso@inescporto.pt

A. Magalhães and M. J. Cardoso are with Faculdade de Medicina, Universidade do Porto, Alameda do Prof. Hernâni Monteiro, 4200-319 Porto, Portugal. amag1976@gmail.com, mjcard@med.up.pt

BCCT.core, without sacrificing the estimated performance.

II. The development of BCCT.core

BCCT.core categorizes the aesthetic result of BCCT into excellent, good, fair, and poor classes. To accomplish the categorization, first, a concise representation of a BCCT image is obtained based on asymmetry, skin colour changes and surgical scar appearance. These measurements are preceded by the semi-automatic localization of fiducial points (nipple complex, breast contour and jugular notch of sternum) on the digital photographs [7], [8]; measures are then supported on these fiducial points. The set of measures is automatically converted onto an overall objective classification of the aesthetical result, using the SVM classifier, trained to predict the overall aesthetical classification on the aforementioned scale of four classes.

A. Features

It is commonly accepted that the cosmetic result after BCCT is mainly determined by visible skin alterations or changes in breast volume or shape. Skin changes can consist of a disturbing surgical scar or radiation-induced pigmentation or telangiectasia. Since there are many ways to describe each of these factors, BCCT.core records all well-known indices and introduced a few more.

The asymmetry is captured using 14 different indices, seven of which require a scale correction between pixels measured on the digital photograph and the length in centimeters on the patient, while the other seven are dimensionless indices, computed as the ratio of two lengths or areas.

The skin colour changes induced by treatment (radiotherapy) were assessed by measuring the dissimilarity of the colour histogram of the two breasts using both the χ^2 and earth movers distance metrics. The dissimilarity was computed both on the 3D histogram and on the histogram of each channel, amounting to 8 different record features.

The surgical scar appearance was translated on a localized colour difference. To compute this, each breast was divided into angular sectors, with the vertex on the nipple; next, the colour histogram for each sector was computed and the similarity between corresponding sectors was measured as for the global colour change; finally, the maximum value of each pair of corresponding sectors was recorded. That amounted for 8 different features [6].

B. The SVM classifier

SVMs have proved themselves as being capable of representing complex classification or mapping functions. They discover the representations using powerful learning algorithms.

In the problem here addressed, there is an inherent ordering between the classes, which motivated the use of models specially targeted for this kind of ordinal data.

Making use of a mapping of the data replication method for ordinal data in SVMs [9], and performing a simplified feature selection, we arrived at an SVM classifier making use of a RBF kernel and requiring only 4 features, two capturing asymmetry, one for skin color changes and one for scar visibility [6].

III. An interpretable model for BCCT

Although the current RBF SVM model presents satisfactory results, its interpretability is not ideal. The RBF kernel is a non-linear kernel, with an implicit mapping to a higher dimensional feature space. The relationships uncovered by the SVM in this implicit feature space are not easily portrayed in the initial space. This nonlinear model is often used as a black box, to which data is presented and the predicted class is outputted. This represents a strong limitation for the caregivers, desiring to understand how the overall classification is being attained.

In this study we consider decision trees and a linear SVM to replace the current RBF SVM. Tree-based models are simple, but widely used, models that work by partitioning the input space into cuboid regions, whose edges are aligned with the axes, and then assigning a simple model (for example, a constant) to each region [10]. A key property of tree based models, which makes them popular in fields such as medical diagnosis, is that they are readily interpretable by humans because they correspond to a sequence of binary decisions applied to the individual input variables. For instance, to predict a patients disease, we might first ask "is their temperature greater than some threshold?". If the answer is yes, then we might next ask "is their blood pressure less than some threshold?". Each leaf of the tree is then associated with a specific diagnosis [10].

Linear models for classification create decision surfaces that are linear functions of the input feature vector and hence are defined by hyperplanes within the input space. The decision surface is therefore a simple weighted sum of the input features, with higher weights being assigned to features more important for the decision process.

A. Study population and a gold standard

This study uses a set of 143 photographs recorded at different breast centers. To train a classifier for performing an automated analysis, a gold standard or ground truth is needed. We use the gold standard taken from the evaluation of patients by an international panel of experts, following a Delphi methodology. A first set of 113 photographs was already available from the initial study [6]; an additional set of 30 photographs was acquired since then. The distribution of the patients over the four different classes is summarized in TABLE I.

TABLE I
DISTRIBUTION OF THE 143 PATIENTS OVER THE FOUR CLASSES.

Class	Excellent	Good	Fair	Poor	**Total**
# cases	20	74	34	15	143

Since the extraction of the features from the image is still not completely automated, we asked 8 users to manually place the necessary fiducial points; all 143×8 cases were later used to develop the classification models.

B. Feature selection

In supervised learning, variable selection is used to find a subset of the available inputs that accurately predict the output. The objective of variable selection is: improve prediction performance of the model, provide faster and more cost-effective predictors, and provide a better understanding of the underlying process that generated the data [11].

Features selection algorithms can be divided into 'filter' and 'wrapper', according to the nature of the methods used to evaluate features. Wrapper algorithms evaluate features with the classification accuracy provided by a target classification algorithm. A target classifier is included in the feature selection process of the wrapper approach. This leads to the improvement of classification accuracy of the wrapper classifier, but it also contributes to the increase of the computation time. Furthermore, because the derived feature subset is biased to the wrapped classification algorithm, good performance may not occur when the feature subset is used to build models with other classification algorithms.

Filter algorithms are independent of any learning algorithms; feature evaluation is carried out according to a measure that evaluates the goodness of individual features, such as correlation. Features are ranked according to their values on the selected measure. The first ranked features are chosen from the entire set of available characteristics, according to prior domain knowledge or a user-specified threshold value.

Evaluating all possible subsets is unfeasible; remember that we recorded 30 features, leading to 2^{30} different subsets of features. Therefore, we performed a simplified selection, also benefiting from our previous work [6]. We had already concluded that the 7 dimensionless asymmetry features perform as good as or even better than the dimensional asymmetry features, with the additional advantage of not requiring the scale mark in the patient photograph [6]. Therefore, 7 asymmetry features were immediately discarded.

For the remaining 23 $(7 + 8 + 8)$ features, the squared correlation coefficient (R^2) was computed. The squared correlation coefficient, also referred as coefficient of determination, is the proportion of target variation explained by a single input variable. The coefficient ranges from 0, when there is no linear relationship between an input and the target, to 1, for an input that explains all of the target variability. Simultaneously, we also analyzed the behaviour of each feature when going from the excellent to the poor class. It is expected that the average value of any asymmetry, colour or scar visibility measure increases monotonically from the excellent class to the poor class. From the features exhibiting a monotonous behavior, we kept the 3 dimensionless asymmetry features, the 2 colour change features and the 2 scar visibility features with highest R^2 coefficient:

1) pLBC: dimensionless difference between levels of inferior breast contour;
2) pBCD: dimensionless difference between lengths of left and right breast contours;
3) pBAD: dimensionless difference between areas of left and right breasts;
4) $c\chi^2a$: distance between the histograms of the left and right breasts measured in the a channel of the CIE L*a*b* colour space, using the χ^2 measure;
5) $cEMDa$: distance between the histograms of the left and right breasts measured in the a channel of the CIE L*a*b* colour space, using the earth movers distance;
6) $s\chi^2a$: visibility of the surgical scar captured in the a channel of the CIE L*a*b* colour space, using the χ^2 measure;
7) $sEMDa$: visibility of the surgical scar captured in the a channel of the CIE L*a*b* colour space, using the earth movers distance.

Detailed information about these features can be found in [6].

Finally, with the remaining 7 features, we considered all possible subsets. The performance of each subset of features was estimated using a four-fold cross-validation scheme.

C. Dealing with Decision Costs

In this application, the default assumption of equal misclassification costs underlying machine learning techniques is violated. Caregivers consider an error in a true excellent or true poor patient more penalizing than an error in the middle classes (fair or good). Moreover, failure to a contiguous class is not as serious as failure to a non contiguous class. From these considerations we defined a cost matrix reflecting the penalty of classifying samples from one class as another:

$$C = \begin{bmatrix} 0 & 2 & 4 & 6 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 6 & 4 & 2 & 0 \end{bmatrix}$$

where $C(i, j)$ is the cost of classifying a point into class j if its true class is i. The cost matrix was taken into consideration during the model building process.

IV. RESULTS

The problem here addressed involves classifying examples into classes which have a natural ordering. Therefore we adopted classification methods specific for this kind of data. The SVM classifier was based on the data replication method for ordinal data [9], both with linear and RBF kernels. We performed a "grid-search" on the parameters of the models using cross-validation in the training set (h and s parameters were left constant at 1 and 2, respectively). The test results concerning the misclassification error are summarized in TABLE II, for the first ranked feature subsets:

Feature Set 1: {pLBC, pBAD, $c\chi^2a$, $s\chi^2a$}
Feature Set 2: {pLBC, pBCD, $c\chi^2a$, $s\chi^2a$}
Feature Set 3: {pLBC, pBCD, pBAD, $c\chi^2a$, $s\chi^2a$}

For the decision tree algorithm, the cross-validation scheme was carried out over the pruning level of the tree. The misclassification error for the test phase is summarized on TABLE III. In both tables, the weighted error was obtained by weighting each prediction error result with the cost of the misclassification.

TABLE IV

CONFUSION MATRICES FOR BEST CLASSIFIERS, IN THE TEST SET.

(a) RBF SVM

Predict \ True	Excellent	Good	Fair	Poor
Excellent	30	18	0	0
Good	10	68	2	0
Fair	6	33	39	2
Poor	0	0	0	32

(b) LINEAR SVM

Predict \ True	Excellent	Good	Fair	Poor
Excellent	36	12	0	0
Good	15	59	6	0
Fair	8	48	24	0
Poor	0	0	0	32

(c) DECISION TREE

Predict \ True	Excellent	Good	Fair	Poor
Excellent	19	26	3	0
Good	6	71	3	0
Fair	1	56	23	0
Poor	0	0	8	24

TABLE II

MISCLASSIFICATION ERROR USING THE SVM CLASSIFIER.

Feature Set	kernel	C	gamma	**Error**	**Weighted Error**
1	RBF	16	0.5	0.42	0.63
2		1	0.5	0.30	0.52
3		1	0.75	0.33	0.56
1	Linear	16	-	0.37	0.52
2		64		0.37	0.58
3		4		0.38	0.60

TABLE III

MISCLASSIFICATION ERROR USING THE DECISION TREE CLASSIFIER.

feature set	# levels	**Error**	**Weighted Error**
1	7	0.47	0.63
2	7	0.43	0.61
3	7	0.46	0.64

A first observation is that the current RBF SVM model performs better than both the linear model and the decision tree. Nevertheless the LINEAR SVM follows closely the accuracy of the nonlinear model, in particular if we focus on the weighted error. The decision trees did not achieve the same level of performance. This is likely related with the small size of the available dataset, insufficient for reliable learning with decision trees. It is also instructive to analyze the confusion matrices of the models on TABLE IV (note that we use data taken from 8 different users).

As observed, there is no confusion between the end classes. Moreover, most of the errors are to a contiguous class and concentrated in the 'good' class, with true 'good's being predicted either as 'excellent' or 'fair'. The elimination of these errors will likely require the integration of new features, capturing more information to distinguish these cases. We should also not forget the likely existence of inconsistencies in the reference information, due to the subjectivity inherent to the process of obtaining them.

V. CONCLUSIONS

We have developed different accurate and interpretable models for the assessment of aesthetic result of BCCT. Having as baseline the current nonlinear RBF SVM model deployed with BCCT.core, we compared the accuracy of a decision tree and a LINEAR SVM, models that facilitate the comprehension of factors with impact on the decision. We have shown that the linear model achieves a performance very similar to the RBF SVM, with the obvious advantages of simplicity and interpretability.

The replacement of the current model in BCCT.core by the linear SVM model will increase the interpretability and acceptance of BCCT.core, enabling caregivers to validate their experimental results, and improving trust on this kind of software for aesthetic evaluation of BCCT.

We now intend to increase the robustness of BCCT.core with the introduction of new features. Namely, we expect that features extracted from patients lateral views, combined with the results presented in this paper, will help to increase even further the accuracy of the BCCT.core software.

REFERENCES

[1] M. J. Cardoso, J. S. Cardoso, A. C. Santos, H. Barros, and M. C. Oliveira, "Interobserver agreement and consensus over the esthetic evaluation of conservative treatment for breast cancer," *The Breast*, vol. 15, pp. 52–57, february 2006.

[2] R. D. Pezner, M. P. Patterson, L. R. Hill, N. Vora, K. R. Desai, J. O. Archambeau, and J. A. Lipsett, "Breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer," *International Journal of Radiation Oncology Biology Physics*, vol. 11, pp. 575–578, 1985.

[3] E. V. Limbergen, E. V. Schueren, and K. . V. Tongelen, "Cosmetic evaluation of breast conserving treatment for mammary cancer. 1. proposal of a quantitative scoring system," *Radiotherapy and Oncology*, vol. 16, pp. 159–167, 1989.

[4] D. R. H. Christie, M.-Y. O'Brien, J. A. Christie, T. Kron, S. A. Ferguson, C. S. Hamilton, and J. W. Denham, "A comparison of methods of cosmetic assessment in breast conservation treatment," *Breast*, vol. 5, pp. 358–367, 1996.

[5] S. K. Al-Ghazal, R. W. Blamey, J. Stewart, and A. L. Morgan, "The cosmetic outcome in early breast cancer treated with breast conservation," *European Journal of Surgical Oncology*, vol. 25, pp. 566–570, 1999.

[6] J. S. Cardoso and M. J. Cardoso, "Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment," *Artificial Intelligence in Medicine*, vol. 40, pp. 115–126, 2007.

[7] J. S. Cardoso, L. F. Teixeira, and M. J. Cardoso, "Automatic breast contour detection in digital photographs," in *Proceedings of the International Conference on Health Informatics (HEALTHINF 2008)*, L. Azevedo and A. Londral, Eds., vol. 2, 2008, pp. 91–98.

[8] J. S. Cardoso, R. Sousa, L. F. Teixeira, and M. J. Cardoso, "Breast contour detection with stable paths," in *Biomedical Engineering Systems and Technologies*, ser. Communications in Computer and Information Science, A. Fred, J. Filipe, and H. Gamboa, Eds., vol. 25. Springer-Verlag Berlin Heidelberg, 2009, pp. 439–452.

[9] J. S. Cardoso and J. F. P. da Costa, "Learning to classify ordinal data: the data replication method," *Journal of Machine Learning Research*, vol. 8, pp. 1393–1429, 2007.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[11] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 1157–1182, 2003.